Proceedings of 2010 International Conference on Systems in Medicine and Biology
16-18 December 2010, IIT Kharagpur, India

Mining Association Rules from HIV-Human Protein Interactions

Anirban Mukhopadhyay*, Ujjwal Maulik†, Sanghamitra Bandyopadhyay‡ and Roland Eils§

*Dept. of Computer Science and Engineering, University of Kalyani, Kalyani-741235, India. Email: anirban@klyuniv.ac.in
†Dept. of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India. Email: umaulik@cse.jdvu.ac.in
‡Machine Intelligence Unit, Indian Statistical Institute, Kolkata-700108, India. Email: sanghami@isical.ac.in
§Dept. of Theoretical Bioinformatics, German Cancer Research Center, Heidelberg-69120, Germany. Email: r.eils@dkfz.de

***Abstract*—Identifying possible viral-host protein-protein interactions is an important and useful approach in developing new drugs targeting those interactions. In this article, a recently published dataset containing records of interactions between a set of HIV-1 proteins and a set of human proteins has been analyzed using association rule mining. The main objective is to identify a set of association rules among the human proteins with high confidence. The well-known *Apriori* algorithm has been utilized for discovering the association rules. Moreover, we have predicted some new viral-human interactions based on the discovered association rules.**

***Index Terms*—Protein-protein interaction, HIV-1–human interaction, association rule mining, Apriori algorithm.**

I. Introduction

Analysis of the regulation between viral and host proteins in different organisms is an important step to uncover the underlying mechanism of various viral diseases. Human immunodeficiency virus (HIV) is a lentivirus (a member of the retrovirus family with long incubation period) that can lead to acquired immunodeficiency syndrome (AIDS), a condition in humans in which the immune system begins to fail, leading to life-threatening infections [1]. HIV-1 is a species of the HIV virus that relies on human host cell proteins in virtually every phase of its life cycle. One of the main goals in research of Protein-Protein Interaction (PPI) is to predict possible viral-host interactions. This is specifically aimed at assisting drug developers targeting protein interactions for the development of specially designed small molecules to inhibit potential HIV-1–human PPIs. Targeting protein-protein interactions has relatively recently been established to be a promising alternative to the conventional approach to drug design [2], [3].

There are several computational approaches for predicting PPIs [4], [5], [6], [7], [8], [9], [10]. Most of these approaches are mainly used for determining PPIs in a single organism, such as yeast, human etc. However, determination of PPIs across multiple organisms such as between viral proteins and the corresponding host proteins can contribute to the development of new therapeutic approaches and design of drugs for these viral diseases. Moreover, these methods are mainly based on designing some classifiers which need both positive and negative samples for PPIs. Although there are several online resources that systematically store information about experimentally validated interacting proteins, there is no such resource for non-interacting proteins which should be used as the ideal negative samples. Therefore in most of the works in this area, negative samples are prepared by taking random protein pairs which are not found in the interaction database. This is done with the expectation that this random protein pairs are less likely to interact physically, which may not be true always. The performance of the classifier highly depends on the choice of the negative samples.

With this observation, in this article we have proposed an approach based on association rule mining [11] that uses information based on positive samples of experimentally validated PPIs only. The PPI information among HIV-1 and human proteins are organized as a binary matrix with rows representing the viral proteins and columns representing the human proteins. Thereafter, the well-known *Apriori* algorithm [11], [12] has been used for discovering association rules among the columns, i.e., human proteins. Finally these rules have been utilized to predict some new viral-host interactions.

The remaining part of the article is organized as follows: the next section discusses the basic concepts of association rule mining and describes the *Apriori* algorithm in brief for association rule mining in this regard. Section III discusses how association rule mining can be used for predicting new interactions based on an existing HIV-1–human PPI database. In Section IV, the experimental results have been provided. Finally Section V concludes the article.

II. Association Rule Mining

The principle of association rule mining (ARM) problem lies in the market basket or transaction data analysis. Many information is hidden in the day to day transactions taking place in supermarkets. For example a customer who is buying nappy also likes to purchase baby food in the same time. Association analysis is the discovery of rules showing attribute–value associations that occur frequently [13]. Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n items and X be an itemset where $X \subseteq I$. A k-itemset is a set of k items. Let $T = \{(t_1, X_1), (t_2, X_2), \ldots, (t_m, X_m)\}$ be a set of m transactions, where t_i and X_i, $i = 1, 2, \ldots, m$, are the transaction identifier and the associated itemset respectively. The *cover* of an itemset X in T is defined as follows:

$$cover(X, T) = \{t_i | (t_i, X_i) \in T, X \subseteq X_i\}. \quad (1)$$

 978-1-61284-038-3/10/$26.00 ©2010 IEEE

The *support* of an itemset X in T is

$$support(X,T) = |cover(X,T)| \tag{2}$$

and the *frequency* of an itemset is

$$frequency(X,T) = \frac{support(X,T)}{|T|}. \tag{3}$$

Thus support of an itemset X is the number of transactions where all the items in X appear in each transaction. The frequency of an itemset is the probability of its occurrence in a transaction in T. An itemset is called frequent if its support in T is greater than some threshold min_sup. The collection of frequent itemsets with respect to a minimum support min_sup in T, denoted by $\mathcal{F}(T, min\_sup)$ is defined as

$$\mathcal{F}(T, min\_sup) = \{X \subseteq I : support(X,T) > min\_sup\}. \tag{4}$$

The objective of ARM is to find all rules of the form $X \Rightarrow Y$, $X \bigcap Y = \emptyset$ with probability $c\%$, indicating that if itemset X occurs in a transaction, the itemset Y also occurs with probability $c\%$. X and Y are called the *antecedent* and *consequent* of the rule respectively. Support of a rule denotes the percentage of transactions in T that contains both X and Y. This is taken to be the probability $P(X \bigcup Y)$. An association rule (AR) is called *frequent* if its support exceeds a minimum value min_sup.

The confidence of a rule $X \Rightarrow Y$ in T denotes the percentage of the transactions in T containing X that also contains Y. It is taken to be the conditional probability $P(X|Y)$. In other words,

$$confidence(X \Rightarrow Y, T) = \frac{support(X \bigcup Y, T)}{support(X,T)}. \tag{5}$$

A rule is called *confident* if its confidence value exceeds a threshold min_conf. Formally the ARM problem can be defined as follows: Find the set of all rules R of the form $X \Rightarrow Y$ such that

$$\begin{aligned} R = \quad & \{X \Rightarrow Y : X, Y \subseteq I, X \bigcap Y = \emptyset, \\ & X \bigcup Y \in \mathcal{F}(T, min\_sup), \\ & confidence(X \Rightarrow Y, T) > min\_conf\}. \end{aligned} \tag{6}$$

Generally the ARM process consists of the following two steps [14], [15] :

1) Find all frequent itemsets.
2) Generate strong ARs from the frequent itemsets.

Apart from the above mentioned general framework adopted in most of the research in ARM, there is another approach for immediately generating a large subset of all ARs [16].

The number of itemsets grows exponentially with the number of items $|I|$. A commonly used algorithm for generating frequent itemsets is the *Apriori* algorithm [11], [12]. This is based on the idea that if even one subset of an itemset X is not frequent, then X cannot be frequent. It starts from all itemsets of size one, and proceeds in a recursive fashion. If any itemset X is not frequent then that branch of the tree is pruned, since any possible superset of X can never be frequent.

III. ARM IN VIRAL-HOST PPI

In this section, the procedure for mining ARs from HIV-1–human PPI network has been described. First we describe the preparation of the input data set. Thereafter, how to apply *Apriori* algorithm on the input data set to discover highly confident rules and how these rules are used to predict new interactions are discussed.

A. Preparation of Input Data Set

The interaction information reported between HIV-1 and human proteins in [17], which has been prepared based on a recently published PPI data set [18] has been collected. The interaction data set contains three types of interactions, viz., group-1 interactions representing direct physical interactions, group-2 interactions representing indirect interactions, and the other interactions that are predicted using their method. The third kind of interactions are marked as 'Novel' interactions which are computationally predicted by Tastan et. al. [17]. We have concentrated our study only on group-1 and group-2 interactions, since these are experimentally validated. There are total 1288 group-1 and group-2 interactions between 17 HIV-1 proteins and 773 human proteins. We have constructed a binary matrix of size 17×773. An entry of 1 in the matrix denotes the presence of interaction between the corresponding pair of HIV-1 and human proteins, and an entry of 0 represents the absence of any information regarding the interaction of the corresponding viral and human proteins. Initially it is treated as non-interaction. The resulting binary matrix is treated as the input to the ARM algorithm.

B. Application of Apriori ARM Algorithm

As discussed above, the rows of the input binary matrix represent the viral proteins and the columns represent the human proteins. Here each row (viral protein) has been considered as a transaction and each column (human protein) has been considered as a item. Thereafter, the *Apriori* algorithm is applied on this transactions to find frequent itemsets and from these frequent itemsets, highly confident ARs are extracted. Note that we have only concentrated on the rules with only one item in the consequent. Hence an example of a discovered AR may be of the form:

$$\{HP_1\ HP_2\ HP_3 \Rightarrow HP_4\}.$$

Here HP_is, $i = 1, 2, 3, 4$, represent four human proteins. In words, the rule can be read as follows: if the human proteins HP_1, HP_2 and HP_3 interact with some viral protein, the human protein HP_4 also interacts with the same viral protein. Corresponding to each rule, there is an associated set of HIV-1 proteins for which the rule is true.

C. Prediction of New Interactions from ARs

We have utilized the discovered ARs that have high confidence to predict new viral-host interactions as follows: Consider a rule $\{HP_1\ HP_2\ HP_3 \Rightarrow HP_4\}$. Suppose in the frequent itemsets, the antecedent of the rule is true for 8 viral proteins $\{VP_1, VP_2, \ldots, VP_8\}$. Now without loss

of generality, assume that among these 8 viral proteins, the consequent of the rule is true for the first 6 viral proteins $\{VP_1, VP_2, \ldots, VP_6\}$. Therefore the rule has a confidence of 75% (6 out of 8), which can be thought as reasonably high. From this we can predict that the human protein HP_4 is likely to interact with the viral proteins VP_7 and VP_8 also and the confidence of this prediction is 75%. Thus two new interactions are predicted ($VP_7 \leftrightarrow HP_4$ and $VP_8 \leftrightarrow HP_4$). This way, from all the high-confident rules, we can predict some new interactions with certain levels of confidence.

IV. Results and Discussion

This section describes the experimental results. First the effects of the parameters min_sup and min_conf on the number of extracted ARs have been investigated. Thereafter, the discovered high-confident rules have been reported for a particular min_sup and min_conf. Finally we have reported the newly predicted interactions among the HIV-1 and human proteins.

A. Effects of min_sup and min_conf

To study the effect of min_sup and min_conf on the number of generated ARs, the min_sup has been varied from 30% to 50% (with step size 5%) and the min_conf has been varied from 65% to 90% (with step size 5%), and the number of ARs generated for all the combinations of min_sup and min_conf have been noted down. Fig. 1 shows the variation of number of rules for the different values of min_sup and min_conf. The number of ARs have varied from 0 (min_sup=50%, min_conf=90%) to 671 (min_sup=30%, min_conf=65%). It is evident from the figure that the number of ARs is large when both min_sup and min_conf are on the lower side, whereas the number of ARs is small when both min_sup and min_conf are high. This is quite obvious, since large min_sup and min_conf impose more constraint on the ARs and thus reducing their numbers.

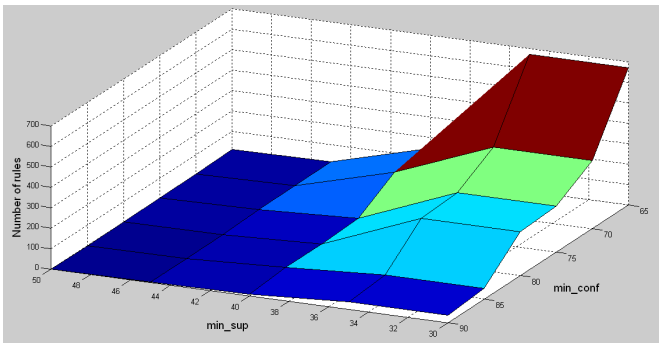

Fig. 1. Variation of number of rules with respect to different values of min_sup and min_conf

B. Demonstration of Generated Rules

In this section, we have reported the rules generated with min_sup and min_conf values of 40% and 70%, respectively. Table I shows these rules with their support and confidence. There are total 34 rules generated with the above mentioned values of min_sup and min_conf. For each rule, the human proteins in the antecedent part and the consequent part of the rule are shown. Note that here we have only considered rules with only one protein in the consequent part. As is evident from the table, there are 21 rules with only one human protein in the antecedent part (Rules 1 to 21), 11 rules with two human proteins in the antecedent part (Rules 22 to 32), and 2 rules with three human proteins in the antecedent part (Rules 33 and 34). As the min_sup value is set to 40%, each rule (the associated frequent itemset) has support>40%. More specifically, there are 22 rules with support 41.2%, 7 rules with support 47.1%, 4 rules with support 52.9% and 1 rule with support 58.8%. Similarly, as min_conf=70%, all rules have confidence>70%. There are 5 rules with confidence 75%, 4 rules with confidence 77.8%, 1 rule with confidence 80%, 15 rules with confidence 85.7%, 1 rule with confidence 87.5%, and 8 rules with confidence 100%.

The rules with one human protein in the antecedent part signify that among all the HIV-1 proteins with which the human protein in the antecedent interacts, the human protein in the consequent is found to interact with c% of those viral proteins, where c is the confidence of the rule. For example, the Rule 1 implies that the human protein ACTG1 interacts with 75% of the viral proteins with which the human protein ACTB interacts. Hence it can be inferred that the human protein ACTG1 also interacts with the remaining 25% viral proteins and the confidence of this inference is 75%. The rules with 2 or 3 human proteins in the antecedent can be interpreted similarly. For example, Rule 22 can be interpreted as follows: the human protein PRKCD interacts with 85.7% of the viral proteins with which both PRKCB1 and PRKCQ interact. Similarly Rule 33 says that the human protein PRKCD interacts with 85.7% of the HIV-1 proteins with which all of PRKCB1, PRKCQ and PRKCA interact.

C. Predicting New Interactions

In this section, we have reported the new interactions predicted from the rules generated (as shown in Table I) using the method discussed in Section III-C. It should be noted that the new interactions can be predicted only from the rules with confidence less than 100%. Hence for predicting new interactions, the rules with 100% confidence are discarded first. Thereafter, the method discussed in Section III-C is applied to predict new HIV-1–human PPIs. Table II reports these new predicted interactions between viral and human proteins. The confidence of the prediction is also given for each predicted interaction. Note that a particular interaction may be predicted from more than one rule, but each of these rules may have different confidence. While reporting the confidence of such predicted interactions, the maximum confidence among the confidences of the predicting rules has been reported. The interaction between a viral protein VP and human protein HP has been denoted by $VP \leftrightarrow HP$. The predicted interactions have been grouped by confidence level. There are unique 22 newly predicted interactions that are

TABLE I
GENERATED RULES FOR min_sup=40% AND min_conf=75%

Rules	Antecedent	Consequent	Support	Confidence
Rule 1	ACTB	ACTG1	47.1	75
Rule 2	ACTB	PRKCA	47.1	87.5
Rule 3	PRKCA	ACTB	52.9	77.8
Rule 4	PRKCB1	PRKCD	41.2	85.7
Rule 5	PRKCQ	PRKCD	41.2	85.7
Rule 6	PRKCB1	PRKCE	41.2	85.7
Rule 7	PRKCQ	PRKCE	41.2	85.7
Rule 8	MAPK3	MAPK1	47.1	100
Rule 9	MAPK1	MAPK3	58.8	80
Rule 10	MAPK3	PRKCA	47.1	75
Rule 11	IFNG	MAPK1	47.1	75
Rule 12	IFNG	PRKCA	47.1	75
Rule 13	CD4	CASP3	41.2	85.7
Rule 14	CASP3	CD4	41.2	85.7
Rule 15	PRKCB1	PRKCQ	41.2	100
Rule 16	PRKCQ	PRKCB1	41.2	100
Rule 17	PRKCB1	PRKCA	41.2	100
Rule 18	PRKCA	PRKCB1	52.9	77.8
Rule 19	PRKCQ	PRKCA	41.2	100
Rule 20	PRKCA	PRKCQ	52.9	77.8
Rule 21	PRKCA	MAPK1	52.9	77.8
Rule 22	PRKCB1, PRKCQ	PRKCD	41.2	85.7
Rule 23	PRKCB1, PRKCA	PRKCD	41.2	85.7
Rule 24	PRKCQ, PRKCA	PRKCD	41.2	85.7
Rule 25	PRKCB1, PRKCQ	PRKCE	41.2	85.7
Rule 26	PRKCB1, PRKCA	PRKCE	41.2	85.7
Rule 27	PRKCQ, PRKCA	PRKCE	41.2	85.7
Rule 28	MAPK3, MAPK1	PRKCA	47.1	75
Rule 29	MAPK1, PRKCA	MAPK3	41.2	85.7
Rule 30	PRKCB1, PRKCQ	PRKCA	41.2	100
Rule 31	PRKCB1, PRKCA	PRKCQ	41.2	100
Rule 32	PRKCQ, PRKCA	PRKCB1	41.2	100
Rule 33	PRKCB1, PRKCQ, PRKCA	PRKCD	41.2	85.7
Rule 34	PRKCB1, PRKCQ, PRKCA	PRKCE	41.2	85.7

shown in Table II. Among them, 7 have 75% confidence, 8 have 77.8% confidence, 1 has 80% confidence, 5 have 85.7% confidence, and 1 has 87.5% confidence.

Unlike classification problem, it is difficult to validate the predicted interactions computationally. Therefore, for validation, we have taken the interactions predicted in [17] (denoted as 'novel' interactions) and compared our predictions with that. Out of 22 predicted interactions, 14 have been found to be predicted in [17] as well. These 14 predicted interactions have been shown in bold faces in Table II. The remaining 8 interactions need more attention and should be studied experimentally to validate the computational predictions.

V. CONCLUSIONS

This article poses the problem of predicting new HIV-1–human protein interactions based on the existing PPI database as an association rule mining problem. This is motivated due to the lack of experimentally validated negative samples needed to pose the problem as a classification problem. The well-known *Apriori* algorithm has been used to mine the frequent itemsets from the HIV-1–human PPI database organized as a binary matrix. The association rules of high confidence have been generated and a novel method for predicting new interactions from the generated rules has been proposed. The proposed method has been shown to predict new viral-host interactions with certain confidence levels. It is needed to validate the predicted interactions both computationally and biologically. The authors are working in these directions.

ACKNOWLEDGEMENT

A part of the work was carried out when Dr. U. Maulik was visiting German Cancer Research Center, Heidelberg, Germany with Humboldt Fellowship for Experienced Researchers.

REFERENCES

[1] A. L. DeFranco, R. Locksley, and M. Robertson, *Immunity: the immune response in infectious and inflammatory disease*. Oxford University Press, U.K., 2007.

[2] J. Huang and S. L. Schreiber, "A yeast genetic system for selecting small molecule inhibitors of protein-protein interactions in nanodroplets," *Proceedings of the National Academy of Sciences, USA*, vol. 94, no. 25, pp. 13 396–13 401, December 1997.

[3] M. R. Arkin and J. A. Wells, "Small-molecule inhibitors of protein-protein interactions: progressing towards the dream." *Nature Reviews Drug Discovery*, vol. 3, no. 4, pp. 301–317, 2004.

[4] A. Panchenko and T. Przytycka, *Protein-protein Interactions and Networks: Identification, Computer Analysis, and Prediction*. London: Springer-Verlag, 2008, vol. 9.

[5] R. Jansen *et al.*, "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, pp. 449–453, 2003.

TABLE II
NEW PREDICTED HIV-1–HUMAN PROTEIN INTERACTIONS

Predicted Interactions	Confidence of Prediction (%)
rev↔ACTG1, tat↔ACTG1, **gag_p6↔PRKCA**, **vif↔PRKCA**, **gag_capsid↔MAPK1, gag_capsid↔PRKCA, vpr↔PRKCA**	75
env_gp120↔ACTB, env_gp160↔ACTB, **gag_matrix↔PRKCB1**, rev↔PRKCB1, **gag_matrix↔PRKCQ**, rev↔PRKCQ, **env_gp41↔MAPK1**, **pol_protease↔MAPK1**	77.8
vpr↔MAPK3	80
nef↔PRKCD, pol_protease↔PRKCE, **env_gp160↔CASP3**, env_gp160↔CD4, env_gp160↔MAPK3	85.7
gag_nucleocapsid↔PRKCA	87.5

[6] N. Lin, B. Wu, R. Jansen, M. Gerstein, and H. Zhao, "Information assessment on predicting protein-protein interactions," *BMC Bioinformatics*, vol. 5, no. 154, 2004.
[7] Y. Yamanishi, J. P. Vert, and M. Kanehisa, "Protein network inference from multiple genomic data: A supervised approach," *Bioinformatics*, vol. 20, no. Suppl 1, pp. i363–i370, 2004.
[8] L. Zhang, S. Wong, O. King, and F. Roth, "Predicting co-complexed protein pairs using genomic and proteomic data integration," *BMC Bioinformatics*, vol. 5, no. 38, April 2004.
[9] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, vol. 21, no. Suppl 1, pp. i38–46, 2005.
[10] Y. Qi, J. Klein-seetharaman, and Z. Bar-joseph, "A mixture of feature experts approach for protein-protein interaction prediction," *BMC Bioinformatics*, vol. 8, no. Suppl 10, 2007.
[11] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data (SIGMOD'93)*. New York, NY, USA: ACM, 1993, pp. 207–216.
[12] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499.
[13] S. Bandyopadhyay, U. Maulik, L. B. Holder, and D. J. Cook, *Advanced Methods for Knowledge Discovery from Complex Data (Advanced Information and Knowledge Processing)*. Springer-Verlag, London, 2005.
[14] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining – a general survey and comparison," *SIGKDD Explorations*, vol. 2, no. 1, pp. 58–64, July 2000.
[15] B. Goethals, "Efficient frequent pattern mining," Ph.D. dissertation, University of Limburg, Belgium, 2002.
[16] G. I. Webb, "Efficient search for association rules," in *Proc. Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. The Association for Computing Machinery, 2000, pp. 99–107.
[17] O. Tastan, Y. Qi, J. Carbonell, and J. Klein-Seetharaman, "Prediction of interactions between HIV-1 and human proteins by information integration," in *Proceedings of the Pacific Symposium on Biocomputing*, 2009, pp. 516–527.
[18] W. Fu, B. E. Sanders-Beer, K. S. Katz, D. R. Maglott, K. D. Pruitt, and R. G. Ptak, "Human immunodeficiency virus type 1, human protein interaction database at NCBI," *Nucleic Acids Research (Database Issue)*, vol. 37, pp. D417–D422, 2009.